Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 8297678, 11 pages
https://doi.org/10.1155/2018/8297678

Research Article

An Adaptive Window Size Selection Method for Differentially Private Data Publishing over Infinite Trajectory Stream

Geonhyoung Jo, Kangsoo Jung, and Seog Park

Computer Engineering Department, Sogang University, Republic of Korea

Correspondence should be addressed to Seog Park; spark@sogang.ac.kr

Received 29 May 2018; Revised 2 September 2018; Accepted 18 September 2018; Published 29 October 2018

Guest Editor: Razi Iqbal

Recently, various services based on user's location are emerging since the development of wireless Internet and sensor technology. VANET (vehicular ad hoc network), in which a large number of vehicles communicate using wireless communication, is also being highlighted as one of the services. VANET collects and analyzes the traffic data periodically to provide the traffic information service. The problem is that traffic data contains user's sensitive location information that can lead to privacy violations. Differential privacy techniques are being used as a de facto standard to prevent such privacy violation caused by data analysis. However, applying differential privacy to traffic data stream which has infinite size over time makes data useless because too much noise is inserted to protect privacy. In order to overcome this limitation, existing researches set a certain range of windows and apply differential privacy to windowed data. However, previous researches have set a fixed window size do not consider a traffic data's property such as road structure and time-based traffic variation. It may lead to insufficient privacy protection and unnecessary data utility degradation. In this paper, we propose an adaptive window size selection method that consider the correlation between road networks and time-based traffic variation to solve a fixed window size problem. And we suggest an adjustable privacy budget allocation technique for corresponding to the adaptive window size selection. We show that the proposed method improves the data utility, while satisfying the equal level of differential privacy as compared with the existing method through experiments that is designed based on real-world road network.

1. Introduction

Today, various services based on user location are emerging as the wireless Internet and sensor technology develop, and VANET (vehicular ad hoc network), which provides wireless communication between vehicles, is also highlighted as one of the services. In the VANET environment (Figure 1), users can know about traffic jams or emergency situations in real time by the communication between vehicles and RSU (Road-Side Unit) mounted on roads, and VANET administrator provides collected traffic data to external LBS providers. An LBS provider can improve the quality of service by analyzing the traffic data received from the VANET administrator.

However, the traffic data sent to untrustworthy LBS providers, some of whom are untrustworthy, contain sensitive information such as the users' home address and individual trajectories. For example, Table 1(a) shows the trajectory of each vehicle collected in the VANET and this data can be aggregated as shown in Table 1(b). v_i means the vehicle's pseudonym and t_j means the timestamp of the vehicle's trajectory data. Each cell in Table 1(a) represents the location at time t_j and each cell in Table 1(b) represents the aggregated traffic data using original trajectory data. In VANET, the administrator only releases aggregated traffic data instead of the original trajectory data for privacy. In spite of this aggregation, if an attacker knows that vehicle v_3 exists on a certain road at times t_1, t_3, and t_5, the attacker can trace v_3 using the route from "Olympic highway > Hongik University > Sogang bridge" with a 2/3 probability using the aggregated traffic data.

To prevent such attacks, VANET administrators should apply appropriate privacy protection techniques to provide traffic data to external LBS provider. Anonymization techniques, including k-anonymity [1] and l-diversity [2], have been studied for the protection of individual trajectory.

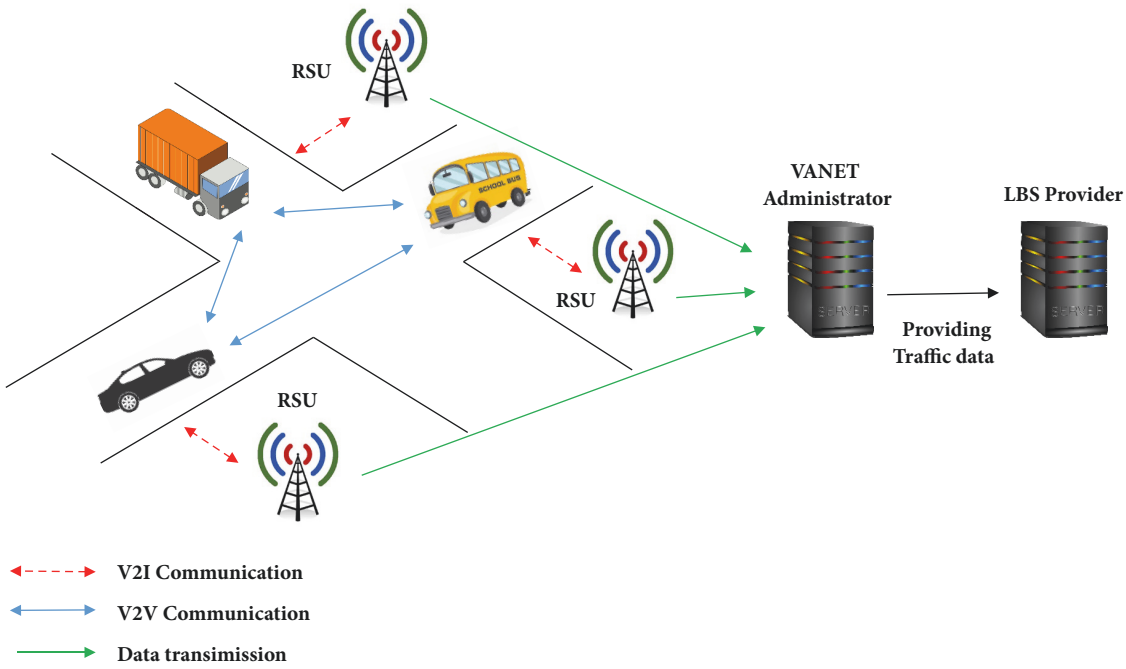


Figure 1: VANET structure and LBS provider.

Table 1

(a) Original trajectories of vehicles

	t_1	t_2	t_3	t_4	t_5	• • •
v_1	Olympic highway				Sogang bridge	• • •
v_2		Sogang bridge			Olympic highway	• • •
v_3	Olympic highway	Teheran street	Hongik university		Sogang bridge	• • •

(b) Aggregated traffic data

	t_1	t_2	t_3	t_4	t_5	• • •
Olympic highway	2	0	0	0	1	• • •
Teheran street	0	1	0	0	0	• • •
Sogang bridge	0	1	0	0	2	• • •
Hongik university	0	0	1	0	0	• • •

However, existing anonymization techniques have a limitation in preventing the inference based on the background knowledge of the attacker while significantly reducing the data usability. By contrast, differential privacy [3], which inserts noise into the data to hide the individual's sensitive information, can prevent the inference of a specific individual trajectory data regardless of the attacker's background knowledge.

A problem with applying differential privacy to periodically provide traffic data is that a privacy budget that determines the amount of noise insertion in differential privacy is divided for each time unit. For example, if we provide traffic data from t_1 to t_{20} in Table 1(b), the total privacy budget is divided by 20 and allocated to each time unit data. As a result, a huge amount of noise is required because the privacy budget is divided into each timestamp in very small pieces, and it significantly deteriorates data utility.

To solve this problem, w-event privacy [4] has been studied. In w-event privacy, a virtual window is assumed to contain w timestamps, and differential privacy is only applied to the windowed data. However, existing w-event techniques inefficiently allocate privacy budgets because the window size is fixed and temporal characteristics (e.g., rush hours) and road network connectivity (e.g., direct route and intersection) are not considered.

In this paper, we show that the existing w-event privacy scheme with a fixed size of w cannot provide sufficient privacy protection and cause unnecessary noise insertion and propose an adaptive window size selection method to overcome this problem. To achieve our goal, we introduce a

method to determine the optimum window size for each road segment through entropy calculation based on road structure and traffic data and propose an improved privacy budget allocation algorithm for adaptive window size selection. The contributions of this work are as follows:

(i) We show that the existing w-event privacy schemes with fixed window sizes cannot prevent the inference of each vehicle's trajectory data and deteriorate the data utility.

(ii) We propose an adaptive window size selection method based on traffic data history and road network structures to solve the fixed window size problem.

(iii) We suggest an adjustable privacy budget allocation to minimize data utility deterioration in adaptive w-event privacy.

The rest of this paper is organized as follows. Section 2 introduces the basic concept of differential privacy and describes the existing w-event privacy technique's concept and limitation. Section 3 suggests an adaptive window size selection method and adjustable privacy budget allocation which prevent an inefficient noise insertion for w-event differential privacy. Section 4 verifies the proposed scheme through experiments, and Section 5 concludes the paper and discusses the future works.

2. Background

2.1. Differential Privacy. Differential privacy is a privacy protection mechanism that prevents private information exposure that is proposed by Dwork in 2006. Dwork proposed a differential privacy to satisfy the requirement that any additional information other than the information obtained from the database itself should not be obtained. For this, Dwork defined a mathematical model to prevent the information exposure which ensures the privacy protection at a specified level ε, which is customized by users. Given two neighboring databases, D_1 and D_2, which differ by only one record, a randomized function K provides ε-differential privacy if all datasets with D_1 and D_2 differ by one element only and all S $\subseteq$ Range(K); i.e.,

$$\frac{\text{Prob}\,(K\,(\text{D}) = S)}{\text{Prob}\left(\text{K}\left(\text{D}'\right) = S\right)} \leq e^{\varepsilon}, \quad S \in \text{Range}\,(K)\,,\ \varepsilon > 0 \qquad (1)$$

This description of differential privacy means that specific individual in the statistical database cannot be deduced correctly by keeping the possibility of a change in query results by inserting/deleting one data to be less than e^{ε}.

According to the definition, the value of ε which is called the privacy budget affects the amount of added noise. As ε decreases, the privacy protection is enhanced. Conversely, as ε increases, the degree of privacy protection decreases.

The most widely used technique for inserting noise to satisfy the differential privacy concept is the Laplace mechanism using the Laplace distribution[]. Let $f(\text{D})$ denote a function of database D. An ε-differentially private Laplace noise mechanism is defined as $\text{L(D)} = f(\text{D})+\text{X}$, where X is a random variable drawn from the Laplace distribution with mean = 0 and standard deviation = $\sqrt{2}\Delta f/\varepsilon$. The Laplace distribution is as follows.

$$\Pr\left(Z\right|\left(\mu, d\right) = \frac{1}{2\text{b}} e^{|x-\mu|/b} \qquad (2)$$

Δf is the sensitivity of the function, which means the maximum value of the change in the query results due to insertion/deletion of a specific individual. That is, the higher the sensitivity and the smaller ε, the greater the probability that a larger noise is inserted.

One of the main properties of differential privacy [5] is that it allows composing of queries. Suppose that the algorithms K_1 and K_2 satisfy ε_1-DP and ε_2-DP, respectively. Then K_1 and K_2 also satisfy the following properties.

(i) Sequential Composition. For any database D, the algorithm that performs $K_1(\text{D})$ and $K_2(\text{D})$ satisfies $(\varepsilon_1 + \varepsilon_2)$-DP.

(ii) Parallel Composition. Let A and B be the partition of any database $\text{D}(\text{A} \cup \text{B} = \text{D}, \text{A} \cap \text{B} = \emptyset)$. Then, the algorithm that performs $K_1(\text{A})$ and $K_2(\text{B})$ satisfies $\max(\varepsilon_1, \varepsilon_2)$-DP.

Because of these two compositions, differential privacy can be applied to complex algorithm implementations.

2.2. Differentially Private Trajectory Data. The privacy protection level in trajectory data depends on how to define the object to be protected. The three types of privacy protection level are as follows.

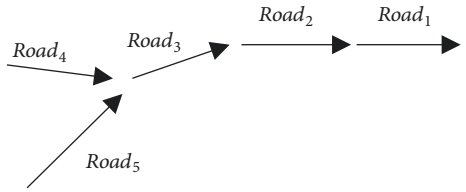
(i) Full Trajectory Privacy. Full trajectory means that the entire trajectory at all timestamps, and it is also called user-level privacy.

(ii) w-Window Trajectory Privacy. w-window trajectory means a partial trajectory from the most recent position of the user to the time before the w time unit, and it is called w-event privacy.

(iii) Event-Level Privacy. Event refers to a single location information at a specific point in time and is called event-level privacy.

To apply differential privacy for traffic data, a privacy budget must be assigned to each timestamp. Therefore, when the user-level privacy technique is applied, the amount of noise at the timestamp t_i becomes exponentially larger than the w-event privacy and event-level privacy. As a result, recent research is being conducted on w-event privacy because w-event privacy can carry out the required level of trajectory analysis without degrading the data usability as much as user-level privacy. Figure 2(b) is the example of w-event privacy as w=3. When traffic data is provided at timestamp t_3, the virtual window w_3 covers t_1, t_2, and t_3 and the vehicle trajectories during this period are protected. Similarly, when traffic data are provided at timestamp t_4, a virtual window w_4 surrounding t_2, t_3, and t_4 is set.

We define w-neighboring as w-event privacy (w-event ε-differential privacy). Furthermore, w-neighboring has the same concept as the neighboring database in differential privacy.



(a)

$t =$	t_1	t_2	t_3	t_4	t_5
R_1	10	20	35	32	29
R_2	20	35	32	29	30
R_3	35	32	29	30	33
⋮	⋮	⋮	⋮	⋮	⋮
R_n	40	51	46	54	48

w_{t_3} w_{t_4} w_{t_5}

(b)

Figure 2: (a) Real trajectories of vehicles and (b) traffic data.

Definition 1 (w-neighboring [4]). For a natural number w, we designate the two trajectories of S_L and S'_L with size L as w-neighboring if their timestamps i_1 and i_2 satisfy the following conditions:

(1) For $1 \le i_1 < i_2 \le L, i_2 - i_1 + 1 \le w$

(2) For $1 \le i_1 < i_2 \le L, S_L[i_1] \ne S'_L[i_1]$ and $S_L[i_2] \ne S'_L[i_2]$

(3) For $1 \le t < i_1, S_L[t] = S'_L[t]$

(4) For $i_2 < t \le L, S_L[t] = S'_L[t]$

Any two trajectories S_L and S'_L are w-neighboring when the different subtrajectories of S_L and S'_L have the maximum size of w.

Definition 2 (w-event ε-differential privacy [4]). If an algorithm K satisfies the equations for all w-neighboring trajectories S and S′, also called w-event ε-differential private algorithm, then

$$\frac{\text{Prob}\,(K\,(S) = R)}{\text{Prob}\,(K\,(S') = R)} \le e^{\varepsilon}, \quad R \in \text{Range}\,(K)\,, \ \varepsilon > 0 \qquad (3)$$

Definitions 1 and 2 also describe how the trajectories of the latest w timestamps can be protected. The privacy level of w-event privacy is determined by ε as well as the general difference privacy. In order to apply the w-event privacy, a privacy budget is allocated for each time unit, and noise is inserted into the traffic data of each time unit. That is, the amount of noise inserted into each time unit increases as the length of w is increased because a privacy budget of ε/w is allocated for each time unit.

2.3. Limitations of Existing w-Event Differential Privacy Technique. The problem with the existing w-event privacy technique [4, 6, 7] is that a fixed-size window is applied to all road segments without considering the correlation among road networks. For example, w = 3 is represented as an ellipse in the traffic data of Figure 2(b), and the trajectories of vehicles move along the roads $R_3 > R_2 > R_1$ in Figure 2(a). At this time, the vehicles at R_3 at any timestamp t_i can be predicted as moving to R_2 at the next timestamp t_{i+1}. In addition, the vehicles in R_2 at timestamp t_i can be predicted as moving to R_1 at the next timestamp. Therefore, an attacker can trace a vehicle trajectory with a high probability when attacks are attempted at t_{i+2}.

On the other hand, the trajectories of vehicles in R_3 at the current timestamp have a relatively lower traceability because of the two entry points of R_3. The above example means that applying the fixed-size window event causes unnecessary data usability degradation and insufficient privacy protection because the possibility of the individual's trajectory inference at a specific time is different according to the feature of the road structure.

In addition to that, the existing w-event privacy technique cannot reflect the variation over time such as rush hour. For example, if no vehicle uses road segment R_5 on special cases, such as road repair work, then there is only one entry point to road segment R_3. Therefore, using a larger window is needed compared to the usual situation.

In this paper, we propose an adaptive window size selection method to solve an existing w-event privacy problem and introduce a privacy budget allocation method suitable for an adaptive window size selection

3. Adaptable w-Event ε-Differential Privacy

The VANET administrator provides a traffic data to the LBS provider while protecting the user's privacy by applying w-event privacy. As mentioned above, the size of w should be determined by reflecting the road structure and time-based traffic variation to satisfy both privacy protection and data utility. In addition to that, the administrator can provide higher data utility by efficiently managing the privacy budget allocated to each time unit within the window. In this paper, we propose an adaptive window size selection based on the entropy value according to the transition probability of each road segment for w-event privacy. In addition, we propose a privacy budget allocation algorithm to overcome the problem of negative or zero privacy budget allocation that may arise from changes in window length.

	time $t-1$		
	R_1	R_2	R_3
t R_1	0.1	0.2	0.7
R_2	0.0	0.0	1.0
R_3	0.3	0.3	0.4

Figure 3: Transition matrix.

3.1. Adaptive Window Size Selection Method. In this section, we explain the algorithm of how to determine the adaptive window size for each road segment at the time t. The size of the window, which is based on the entropy value, is calculated according to the transition matrix at timestamp t. We assume that window size w_i of road segment R_j is an integer ranging from 2 to MaxW (MaxW is an integer). MaxW means the maximum window length which is set by the VANET administrator. Since the VANET administrator periodically collects the identification number and location of all vehicles, it is possible to know which road the vehicle at the road segment R_i at time t was on time t-1. We can calculate the probability using this information and it is referred to as backward transition probability. This transition probability can be expressed as a matrix and we call it a transition matrix (or transition probability matrix). These transition matrices are used to calculate entropy values that can be used to quantify traceability. Each row and column of the transition matrix represent one road segment. The expression of transition matrix tm_t at time t is shown as an example in Figure 3. The transition matrix element tm_t $[R_1][R_3] = 0.7$ means that when a certain vehicle on R_1 is at time t, the probability that it is was on R_3 at time t-1 is 0.7.

The transition probability of the road segment R_j at the time t-1 that can enter the road segment R_i at time t is calculated as follows.

$$\Pr\left(R_j \text{ at } t-1 \mid R_i \text{ at } t\right) = p_j = \frac{traffic_j}{\sum_{p=1}^{n} traffic_p} \quad (4)$$

In this case, $traffic_j$ means the amount of traffic in the road section R_j at that point in time t-1, and n means the number of all possible road segments R_j that can enter R_i including R_i.

The size of window w_i in this selection method for a road segment R_i is as follows. First, the transition matrix tm_t is generated at the time t and let p_j be the probability of each path j that can reach R_i in tm_t. We calculate the entropy value E [8, 9] by using the transition probability.

$$E = -\sum_j \left(p_j \cdot \log p_j\right), \quad (5)$$

p_j = probability of using the j^{th} path

If the entropy E is larger than the threshold θ which is set by VANET administrator for entropy, then the traceability of the vehicle entering R_i is sufficiently low. If the value of E is smaller than θ, then the traceability of the vehicle on R_i from t-1 to t is still high. In this case, window size increases to t-2 and then recalculates the entropy E. This process is repeated until E becomes larger than θ or the window size becomes t-MaxW+1. This process is performed for all road segments. Algorithm 1 obtains the size w_i of the window allocated to each road segment R_i using the above process.

In terms of data utility, the proposed technique guarantees higher data utility than existing w-event technique because the proposed technique determines the window size by calculating the traceability of the vehicle's trajectory in R_i at the time t. If the MaxW is set equal to the fixed window size of the existing scheme, then the proposed technique provides the same level of data utility as the existing scheme in the worst case. This means that the proposed scheme provides better data utility for road segments with a window size less than MaxW.

In terms of data privacy, the proposed window size selection method satisfies MaxW-event ε-differential privacy. We demonstrate the following theorem using a parallel configuration.

Theorem 3. *If the fixed-size window-based technique satisfies MaxW-event ε-differential privacy and uses the same privacy budget allocation algorithm as the proposed technique, then the proposed adaptive-size window-based technique also satisfies MaxW-event ε-differential privacy.*

Proof. Let $R = \{R_1, R_2, ..., R_n\}$ be the entire set of road segments. The vehicles in each road segment cannot be in another road segment at the same time. In other words, the traffic data for each road segment are independent and R is divided into $\{R_i\}$. Applying the adaptive window-based proposed method to the traffic data of each road segment R_i satisfies w_i-event ε-differential privacy. For the time units in the window allocated to R_i, using the same privacy budget allocation algorithm as the existing scheme is assumed. When the results of applying the proposed technique to the traffic data of each road segment by the parallel composition are aggregated, the summation satisfies $\max_i(w_i)$-event ε-differential privacy. At this time, the proposed method satisfies the MaxW-event ε-differential privacy because $\max_i(w_i)$ = MaxW □

3.2. Adjustable Privacy Budget Allocation Algorithm. In the proposed technique, using the existing privacy budget allocation algorithm causes a problem in which a negative- or zero-value privacy budget is allocated because the window size selection is adaptive. In this section, we show a privacy budget allocation problem using the proposed window size selection method and suggest an adjustable privacy budget allocation algorithm to solve the above problem.

3.2.1. Existing Privacy Budget Allocation Algorithm and Limitation. A previous study [4] proposed the w-event ε-differential privacy and budget distribution (BD) and budget

INPUT:
Transition Matrices $tm_1, tm_2, \cdots, tm_t$,
Timestamp t,
Entropy Threshold θ
Maximum Window Size $MaxW$
1: For each road R_i do
2: Clear PATH_QUEUE and PROB_QUEUE
3: PATH_QUEUE.ENQ(R_i) and PROB_QUEUE.ENQ(1.0)
4: For $k = t$ to $t - MaxW + 1$ do
5: Set E = 0
6: $Last\_R$ = PATH_QUEUE.DEQ()
7: For each road R_j incident with $Last\_R$ do
8: Set $p = tm_k[R_i][R_j] \times$ PROB_QUEUE.DEQ()
9: Set E += $-p \times \log(p)$
10: if E > θ then go_to line 12.
11: PATH_QUEUE.ENQ(R_j)
12: PROB_QUEUE.ENQ(p)
13: Set $w_i = t - k + 2$
14: Return $\{(R_1, w_1), (R_2, w_2), \cdots, (R_n, w_n)\}$

Algorithm 1: GetWindowSize.

absorption (BA) algorithms to assign privacy budgets to the timestamp within a window. For window W_i (fixed size MaxW) assigned to each road segment R_i at timestamp t, each algorithm is performed with the following steps:

(i) Similarity Calculation Mechanism. In the similarity calculation step, the noisy traffic data (timestamp l) is compared with raw traffic data at the current timestamp t. The mean absolute error is used for comparison measurement. If the noisy traffic data at timestamp l is similar to the raw traffic data at the current timestamp, then use the noisy traffic data at timestamp l to save on privacy budget [4].

In this paper, as well as the previous techniques, half of the total privacy budget ($\varepsilon/2$) is allocated to this privacy budget in the similarity calculation mechanism, whereas the remaining $\varepsilon/2$ privacy budget is assigned to the privacy budget allocation algorithm for the timestamp within the window. The reason for dividing the privacy budget in the existing technique is that the error in the similarity calculation mechanism affects the entire error value. Hence, the possible maximal privacy budgets are allocated to insert noise into the error value.

(ii) BD Algorithm. The privacy budget ε_t allocated to timestamp t is defined by the following equation.

$$\epsilon_t = \frac{\left(\epsilon/2 - \sum_{k=t-MaxW+1}^{k=t-1} \epsilon_k\right)}{2} \quad (6)$$

The equation implies that only half of the remaining privacy budget in the window is allocated to ε_t (i.e., total privacy budget $\varepsilon/2$ is assigned to the window), which also means an exponential decrease in privacy budget allocated over time and the reuse of the remaining privacy budget allocated to previous timestamps. However, using the BD algorithm is problematic when the window size is very large; namely, the privacy budget is close to zero and the noise inserted into the raw traffic data increases exponentially.

(iii) BA Algorithm. This algorithm starts with an equal privacy budget ($\varepsilon/2 \cdot$MaxW) assignment for each timestamp within the window. If the privacy budget is not used by the similarity calculation mechanism at timestamp t, then the privacy budget is absorbed and used at the next timestamp t+1. Unlike the BD algorithm, the BA algorithm is advantageous in that a certain amount of privacy budget can be allocated even if the privacy budget saving in the similarity calculation mechanism is small.

The existing privacy budget allocation algorithm described above assumes a fixed-size window. If the existing privacy budget allocation algorithm is used for the adaptive-size window technique, then a zero- or negative-value privacy budget is allocated. For example, in the BD algorithm, assume that MaxW is maintained at 2 in four timestamps (t_1–t_4) and increased to MaxW = 5 at timestamp t_5, as shown in Figure 4. First, the privacy budget allocated to timestamp t_1 is $\varepsilon/2$, which is half of total privacy budget ε minus the privacy budget used at the previous timestamp (=0) because the size of the window is 2. Then, the privacy budget allocated to timestamp t_2 is $\varepsilon/4$, which is half of the total privacy budget ε minus the privacy budget ($\varepsilon/2$) used at the previous timestamp. Similarly, $3\varepsilon/8$ and $5\varepsilon/16$ are allocated as privacy budgets at t_3 and t_4, respectively. However, the privacy budget allocated to timestamp t_5 when the size of the window increases to 5 is half of the total privacy budget minus the sum of the privacy budgets allocated to the previous four timestamps. In this case, given that the sum of the privacy budgets allocated to the previous four timestamps is larger than ε, a problem of a negative privacy budget allocation occurs. Similarly, the BA algorithm has a problem of assigning a privacy budget of zero.

3.2.2. Adjustable Privacy Budget Allocation Algorithm. In this section, we propose a new privacy budget allocation algorithm by combining two existing privacy budget allocation

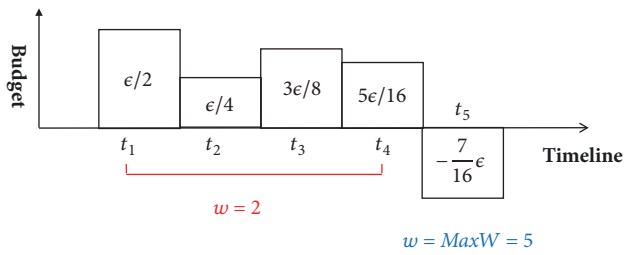


FIGURE 4: Problem examples with the existing privacy budget allocation algorithm.

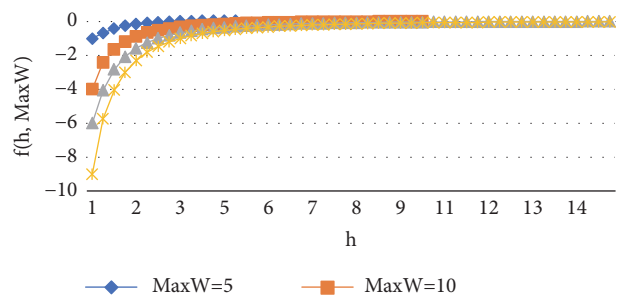


FIGURE 5: Graph of f(h,MaxW).

algorithms to prevent the zero- or negative-value privacy budget allocation in the adaptive-size window, as discussed in Section 3.2.1.

In the proposed technique, the decaying factor h in the BD algorithm is determined as a function of MaxW and not as a fixed constant 2, to solve the problem of negative privacy budget allocation. The privacy budget should always have a positive value even in the worst case. The worst case scenario of privacy budget allocation is the case wherein a window with a size of 2 moves from t = 1 to t = MaxW-1 and increases to MaxW at t = MaxW because the sum of the privacy budgets of ε_t allocated during the time of $1 \le t \le$ MaxW-1 has the largest values. Assuming that the privacy budget allocated to the BD algorithm is ε, the following relations are established to obtain h for this worst case.

First, since no privacy budget is previously used, a privacy budget of ε/h is allocated to timestamp $t = t_1$. Let this value be a_1 ($a_1 = \varepsilon/h$). Then, the privacy budget allocated to timestamp $t = t_2$ is $(\epsilon - \epsilon/h)/h = (\epsilon - a_1)/h$ because the window size is 2, which we designate as a_2. Similarly, the privacy budget allocated to timestamp $t = t_3$ is $(\epsilon - a_2)/h$. This process is repeated until $t = t_{(\text{MaxW-1})}$. When this relation is solved, $a_n = \epsilon/h(h+1)\cdot(-1/h)^{n-1} + \epsilon/(h+1)$ for $1 \le n < MaxW$. Therefore, the privacy budget allocated to timestamp $t = t_{\text{MaxW}}$ is

$$\frac{\epsilon - \sum_{k=1}^{MaxW-1} a_k}{h} = \frac{h^2 + (3 - MaxW)\,h + 1 - MaxW + (-1/h)^{MaxW-1}}{h\,(h+1)^2}\epsilon. \quad (7)$$

Theorem 4. *For $1 \le n < MaxW$, the privacy budget a_n allocated to n^{th} timestamp is $\epsilon/h(h+1)\cdot(-1/h)^{n-1} + \epsilon/(h+1)$.*

Proof. In the equation in Section 3.2.2, $a_n = (\epsilon - a_{n-1})/h = -(1/h)a_{n-1} + \epsilon/h$. Modifying this equation leads to $a_n - \epsilon/(h+1) = -(1/h)(a_{n-1} - \epsilon/(h+1))$. Let $F_n = a_n - \epsilon/(h+1)$, in which the numerical progression $\{F_n\}$ is equivalent to the geometric progression in which geometric the ration is $-1/h$. In this equation, the general terms are $F_n = F_1 \cdot (-1/h)^{n-1}$ and $F_1 = a_1 - \epsilon/(h+1)$. Then, $a_1 = \epsilon/h$ and $F_n = \epsilon/h(h+1)\cdot(-1/h)^{n-1}$. Thus, $a_n = F_n + \epsilon/(h+1) = \epsilon/h(h+1)\cdot(-1/h)^{n-1} + \epsilon/(h+1)$. □

Theorem 5. *The privacy budget allocated to timestamp t_{MaxW} is*

$$\frac{h^2 + (3 - MaxW)\,h + 1 - MaxW + (-1/h)^{MaxW-1}}{h\,(h+1)^2}\epsilon. \quad (8)$$

Proof. The window size is MaxW at time $t = t_{MaxW}$. Thus, the privacy budget at $t = t_{MaxW}$ is $(\epsilon - \sum_{k=1}^{MaxW-1} a_k)/h$. $(\epsilon - \sum_{k=1}^{MaxW-1} a_k)/h$ and $\sum_{k=1}^{MaxW-1} a_k = \sum_{k=1}^{MaxW-1}\{\epsilon/h(h+1)\cdot(-1/h)^{k-1} + \epsilon/(h+1)\}$. After calculation,

$$\begin{aligned}\sum_{k=1}^{MaxW-1} a_k &= \frac{\epsilon}{h\,(h+1)} \sum_{k=1}^{MaxW-1} \left\{\left(-\frac{1}{h}\right)^{k-1} + h\right\} \\ &= \frac{\epsilon}{h\,(h+1)} \left\{\frac{1 - (-1/h)^{MaxW-1}}{1 + 1/h} + h\cdot(MaxW - 1)\right\} \\ &= \frac{\epsilon}{h+1} \left\{\frac{1 - (-1/h)^{MaxW-1}}{h+1} + MaxW - 1\right\}.\end{aligned} \quad (9)$$

Thus,

$$\begin{aligned}&\frac{\epsilon - \sum_{k=1}^{MaxW-1} a_k}{h} \\ &= \frac{h^2 + (3 - MaxW)\,h + 1 - MaxW + (-1/h)^{MaxW-1}}{h\,(h+1)^2}\epsilon\end{aligned} \quad (10)$$

Let $f(h, MaxW) = (h^2 + (3 - MaxW)h + 1 - MaxW + (-1/h)^{MaxW-1})/h(h+1)^2$. Given that ε is a positive real number, the value does not affect the sign of f(h, MaxW). Therefore, assuming that $\varepsilon = 1$, then f(h, MaxW) is the ratio of the privacy budget allocated to $t = t_{\text{MaxW}}$. However, the definition of f(h, MaxW) according to MaxW is an increasing function, as shown in 1 < h < MaxW of Figure 5. Additionally, f(1, MaxW) is always negative for MaxW ≥ 2, while f(MaxW-1, MaxW) is always positive for MaxW ≥ 2. Thus, for a given MaxW ≥ 2, we perform a binary search on $h \in (1, MaxW - 1)$ to find h, such that f(h, MaxW) = 0. □

Algorithm 2 finds and returns the appropriate decaying factor h to prevent negative privacy budget allocation in the worst case.

As mentioned earlier, $\varepsilon/4$ is assigned to the BD algorithm and $\varepsilon/4$ is assigned to the BA algorithm in the proposed technique. The appropriate h derived from the above technique prevents the negative privacy budget from being allocated in the BD algorithm. At this time, a privacy budget that is very close to zero is allocated to $t = t_{\text{MaxW}}$. In the worst case, this condition implies large-scale noise insertion.

```
INPUT:
    Maximum Window Size MaxW
OUTPUT:
    Decaying Factor h
(1)   Set f(h, MaxW) = (h^2 + (3 − MaxW)h + 1 − MaxW + (−1/h)^(MaxW−1))/h(h + 1)^2
(2)   Set Left = 1 and Right = MaxW − 1
(3)   Set h = MaxW − 1
(4)   while Left < Right do
(5)       Set mid = (Left + Right)/2
(6)       if f(mid, MaxW) > 0 then
(7)           if h > mid then
(8)               Set h = mid and Right = mid
(9)       else then
(10)          Left = mid
(11)  Return h
```

Algorithm 2: FindDecayingFactor.

To solve the above problem, the proposed technique allocates the minimum privacy budget to each timestamp using the BA algorithm, as shown in Figure 6(b). Meanwhile, Figure 6(a) shows the probable worst case when using the BD algorithm only with the appropriate decaying factor h for MaxW = 5. In this case, a privacy budget close to zero is assigned over time and a large amount of noise is inserted into the traffic data at the timestamp. However, the proposed algorithm has a privacy budget that is essentially provided by the BA algorithm (i.e., see area represented by the rectangle in Figure 6(b)). Using this case, even if a privacy budget problem close to zero arises from the BD algorithm, a certain amount of privacy budget ($\epsilon/(4 \cdot MaxW)$) is allocated and relatively lesser noise is inserted.

Algorithm 3 provides traffic data to the LBS provider by applying the proposed privacy budget allocation technique to the window allocated to a single road segment R_i.

3.3. Example of Proposed Privacy Budget Allocation Algorithm. In Figure 4, we show an example of the proposed privacy budget allocation algorithm for the timestamps within window W_i assigned to road segment R_i. The novelty of the proposed technique is highlighted in Figure 7.

First, we assume that the sizes of w_i are changed to w_i = MaxW = 5 in $t = t_{MaxW}$ and $w_i = 2$ in $t \in [t_1, t_4]$, as previously shown in Figure 4. Then, we assume that the privacy budget is not saved by the similarity calculation mechanism; i.e., the total privacy budget allocated to the window is $\varepsilon = 1.0$, $\varepsilon/2$ for the similarity calculation mechanism, and $\varepsilon/4$ privacy budget for each of the BA and BD algorithms.

In the case of MaxW = 5, the h causes the privacy budget allocated to $t = t_{MaxW}$ to defer from obtaining a negative value in the BD algorithm, and this value is 3.23402289 according to Algorithm 2. At this time, the privacy budget allocated to $t = t_{MaxW}$ is a value close to zero ($f(h \approx 3.2340, MaxW = 5) \approx 1.22 \times 10^{-18}$). The privacy budget of the proposed technique in the BD algorithm at $t = t_1$ is (ε/4-0)/h=0.07730, and the privacy budget of the proposed technique in the BA algorithm is $\varepsilon/(4 \cdot \text{MaxW}) = 0.05$. Therefore, the privacy budget ε_{t1} allocated to time $t = t_1$ is 0.07730+0.05 = 0.12730. The privacy budget of the BD algorithm allocated to $t = t_2$ is $(\epsilon/4 - 0.07730)/h = 0.05340$ and the privacy budget of the BA algorithm is 0.05. Therefore, $\epsilon_{t_2} = 0.10340$. We can similarly obtain $\epsilon_{t_3} = 0.11079$, $\epsilon_{t_4} = 0.10850$. However, the privacy budget of the BD algorithm allocated to $t = t_{MaxW} = t_5$ is $\{\epsilon/4 - (0.07730 + 0.05340 + 0.06079 + 0.05850)\}/h = 1.1242 \times 10^{-14}$. The privacy budget of the BA algorithm is 0.05, and thus, $\varepsilon_{t5} \approx 0.05$.

Let us consider the case wherein the privacy budget of $\varepsilon/2$ is allocated to the BD algorithm without using the BA algorithm. The privacy budgets allocated to each timestamp using the BD algorithm are $\epsilon_{t_1} = 0.154606$, $\epsilon_{t_2} = 0.10680$, $\epsilon_{t_3} = 0.121582$, $\epsilon_{t_4} = 0.117011$, $\epsilon_{t_5} = 2.248 \times 10^{-14}$. Compared with the proposed technique, this approach is more advantageous when using the BD algorithm only until $t [t_1, t_4]$. However, if the BD algorithm is used only at $t = t_5$, the privacy budget becomes very small and a large amount of noise (with absolute size of $10^{13} \sim 10^{14}$) is inserted. In addition, proposed technique dramatically deteriorates data utility because the magnitude of this noise is inserted into the traffic volume of a single road segment. Figure 4 shows the privacy budget allocated to each timestamp when the proposed technique is applied. As shown by Figure 7, the proposed technique can solve the problem of the existing technique (i.e., see Figure 4).

4. Experimental Results

4.1. Experimental Environment. The dataset used in the experiment is a set of synthetic data based on the actual road network provided by the Seoul Metropolitan Transportation Information System [10]. The traffic dataset and average speed data are generated based on the traffic volume data for each road segment in Seoul city.

Seoul city provides a 214 road segments' traffic volume data and 4751 road segment's average speed data. In this experiment, we generate 4751 road segments' traffic volume data based on the actual traffic data distribution of 214 road

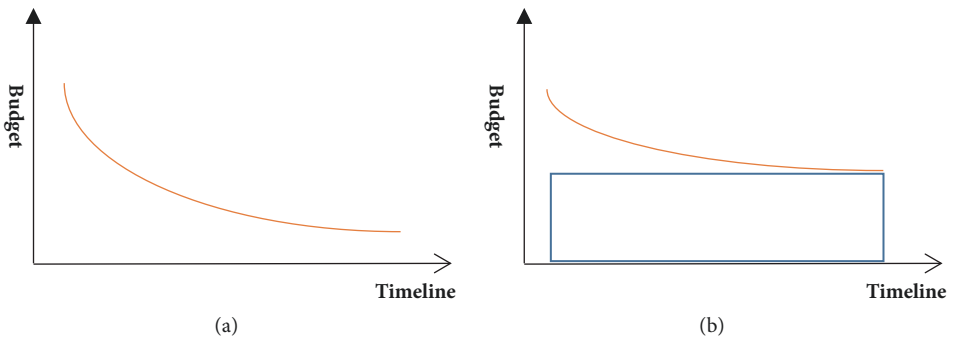


Figure 6: (a) BD algorithm and (b) proposed algorithm.

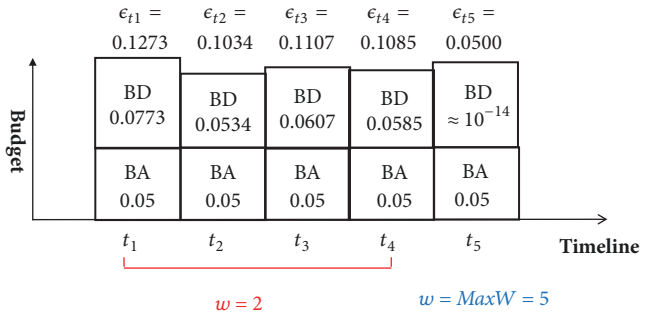


Figure 7: Example of proposed privacy budget allocation algorithm.

segments. Among the 4751 road segments, 456 roads have a straight segment, and the number of generated vehicles is 50,000–100,000 per time unit. We assume that the data on traffic volume are provided every 5 minutes to allow the updating of the optimal route in 5-minute cycles in the navigation service of the domestic LBS service, such as T-map [11], which is the standard for collecting the locational information of vehicles (i.e., within 5 minutes in the VANET environment). Therefore, the whole number of time unit is 288, which can be divided by 5 minutes every day. Moreover, t = 1 is assumed to be the time at midnight (00:00 H). Finally, we assume that the time required for vehicles to travel is from 1 hour to 1 hour 30 minutes, and the MaxW range is defined as 12–18 timestamps in this experiment.

4.2. Experimental Method and Evaluation Criteria. We compared the proposed technique with the BD and BA algorithms in fixed-size window conditions [4]. We estimated mean absolute error (MAE) and mean relative error (MRE) to evaluate the amount of noise inserted into each road segment. The results show the average values obtained by repeating the experiment 100 times.

$$\text{MAE} = \sum_{r \in R} \left(\frac{\left| D(r) - D'(r) \right|}{|R|} \right) \tag{11}$$

$$\text{MRE} = \sum_{r \in R} \left(\frac{\left| D(r) - D'(r) \right|}{|R|} * D(r) \right) \tag{12}$$

where (D(r) is the original traffic volume for the road segment, D′(r) is the noisy traffic data, and R is the set of all road segments.

4.3. Experiment Result about a Threshold Value θ for Entropy. Threshold θ determines the window size allocated to the road segment. As the value of θ increases, the average window size increases using the proposed GetWindowSize algorithm. This finding implies a reduction of the privacy budget allocated to each time unit and the insertion of additional noise.

Figures 8 and 9 show the MAEs and MREs of the proposed, BD, and BA algorithms when MaxW = 15. In the case of the BD and BA algorithms with fixed-size windows, almost no difference is observed between MAE and MRE despite the varying θ. The results show that both algorithms set the fixed-size windows MaxW = 15 in all road segments. However, in the case of the proposed technique, the MAE and MRE tended to increase because the average window size increased as θ increased. Consequently, the data utility of the proposed algorithm is better than the existing algorithms.

4.4. Experiment Result about a Maximum Window Length (MaxW). MaxW is defined as the maximum value of the window size assigned to each road segment. As MaxW increased, the average window size increased when using the proposed window allocation algorithm of GetWindowSize. This means that the privacy budget allocated to each time unit is reduced and the insertion of noise is increased.

Figures 10 and 11 show the MAEs and MREs of the proposed, BD, and BA algorithms at $\theta = 0$ according to MaxW. In the case of BD and BA algorithms with fixed-size windows, both MAEs and MREs increased as MaxW increased because both algorithms allocate a window with a size of MaxW in all road segments. By contrast, the proposed technique's average window size is determined by θ. Given that only a part of the road segment with entropy $E \leq \theta$ is affected by MaxW, the MAEs and MREs tended to increase relatively slow. The data utility of the proposed algorithm is better than the existing algorithms.

4.5. Experiment Result about a GetWindowSize Algorithm Processing Time. Unlike the existing method, the proposed method determines the window size by calculating the transition probability of each road segment. Therefore, the

INPUT:
Desired Privacy Leakage ϵ,
Timestamp t,
Road R_i,
Traffic Volume at timestamp t for all road D_t
Window size w_i,
Maximum Window Size $MaxW$
OUTPUT:
Noisy Traffic Volume of R_i $D'_t[R_i]$ or $D_\ell[R_i]$
(1) Let ℓ be the latest timestamp that releases noisy traffic volume
(2) Let D_ℓ be latest released noisy traffic volume
(3) Set $dis = |D_\ell[R_i] - D_t[R_i]|$ and $\lambda_{dis} = (2 \cdot MaxW)/\epsilon$
(4) Set $dis = dis + \text{Lap}(\lambda_{dis})$
(5) Set $h = \text{FindDecayingFactor}(MaxW)$
(6) Set $\epsilon_{t,\text{BD}} = (\epsilon/4 - \sum_{k=t-w_i+1}^{t-1} \epsilon_{k,\text{BD}})/h$
(7) Set to_nullify $= \epsilon_{l,\text{BA}}/(\epsilon/(4 \cdot MaxW)) - 1$
(8) if $t - \ell \leq$ to_nulify then
(9) Return $D_\ell[R_i]$ // Enforce to skip publication $D_t[R_i]$
(10) else
(11) Set to_absorb $= t - (l+$to_nullify$)$
(12) Set $\epsilon_{t,\text{BA}} = \epsilon/(4 \cdot MaxW) \cdot \min(\text{to\_absorb}, w_i)$
(13) and $\lambda_{Budget} = 1/(\epsilon_{t,\text{BD}} + \epsilon_{t,\text{BA}})$
(14) if $dis > \lambda_{Budget}$ then
(15) Return $D'_t[R_i] = D_t[R_i] + \text{Lap}(\lambda_{Budget})$
(16) else
(17) Return $D_\ell[R_i]$

Algorithm 3: Publication.

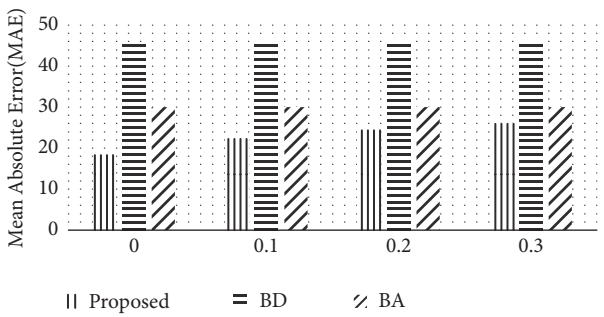


Figure 8: MAE of the proposed and existing methods with varying θ.

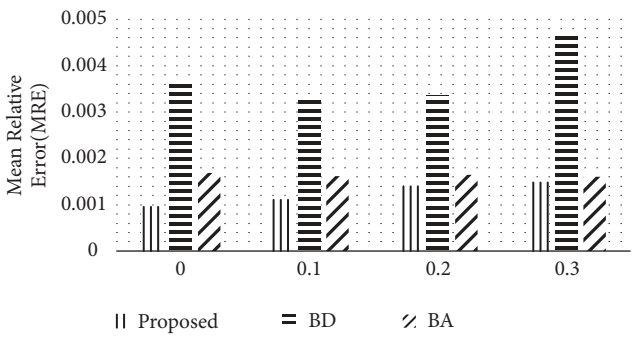


Figure 9: MRE of the proposed and existing methods with varying θ.

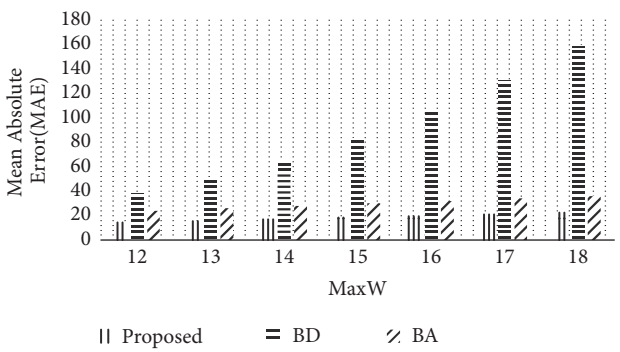


Figure 10: MAE of proposed and existing methods with varying $MaxW$.

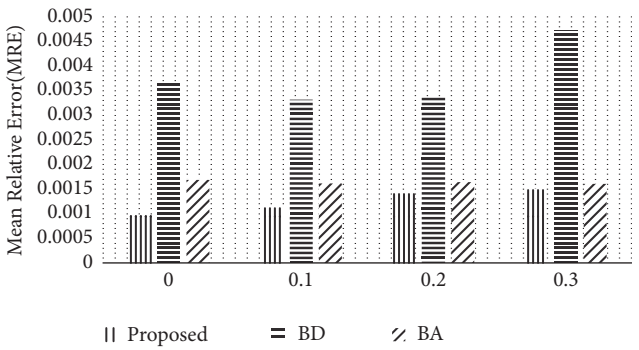


Figure 11: MRE of proposed and existing methods with varying $MaxW$.

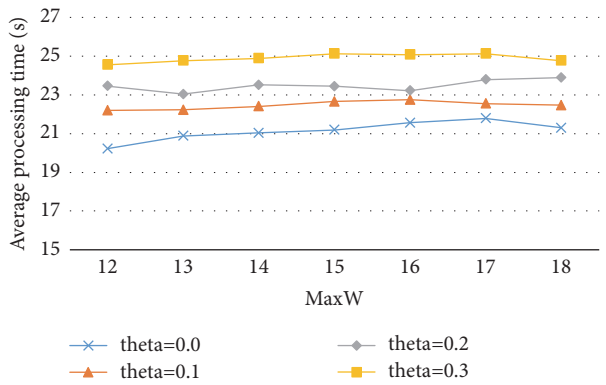


Figure 12: Average processing time of GetWindowSize algorithm.

execution time of GetWindowSize is an extra cost to be paid in the proposed technique compared to the existing technique. Figure 12 shows the average time taken to find the length of the window of all road sections during one time unit, according to the changes of θ and MaxW. Experimental results indicate that the proposed method is not problematic because it is performed within 5 minutes which is the general data publication cycle.

As shown in the experimental result, it can be seen that, as the threshold θ increases, the length of the road segment to be calculated increases, so that the execution time increases. However, it can be confirmed that the increases are not significant and linearly increase. MaxW does not have a significant effect on the results, as the experimental result shows that the road segment reaching MaxW occupies only a part of the entire road segment.

5. Conclusion

Massive amounts of data can now be collected by mobile devices, sensors, and Web services. These large amounts of data represent data stream properties, such as real-time data utilization by users. However, individual sensitive information contained in the collected data may be exposed by the data analysis. Therefore, the data provider must be able to apply the appropriate privacy protection techniques to provide data utility to users. To this end, we propose an adaptive window size selection method that considers the correlation among road segments and time specificity that were not covered by the existing w-event privacy technique for traffic data. In addition, we propose a privacy budget allocation algorithm to overcome the zero or negative privacy budget allocation problem that can occur while using adaptive-size windows. The proposed technique shows better data utility than the existing techniques on the basis of experimental results. Future work may include research on finding the correlations that can occur not only in traffic volume data but also in data in other domains and developing algorithms that can allocate privacy budgets more efficiently by predicting window sizes.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) (no. 2017-0-00498, A Development of De-Identification Technique Based on Differential privacy).

References

[1] L. Sweeney, "K-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, Atlanta, GA, USA, April 2006.

[3] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, pp. 1–12, 2006.

[4] G. Kellaris, S. Papadopoulos, X. Xiao, and D. Papadias, "Differentially private event sequences over infinite streams," *VLDB Endowment*, vol. 7, no. 12, pp. 1155–1166, 2014.

[5] F. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proceedings of the International Conference on Management of Data and 28th Symposium on Principles of Database Systems (SIGMOD-PODS '09)*, pp. 19–30, July 2009.

[6] Y. Chen, A. MacHanavajjhala, M. Hay, and G. Miklau, "PeGaSus: Data-Adaptive differentially private stream processing," in *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, pp. 1375–1388, November 2017.

[7] "Naver," https://www.naver.com/.

[8] B. Palanisamy and L. Liu, "MobiMix: protecting location privacy with mix-zones over road networks," in *Proceedings of the IEEE 27th International Conference on Data Engineering*, pp. 494–505, Hannover, Germany, April 2011.

[9] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 972–980, Orlando, Fla, USA, March 2012.

[10] "Seoul Metropolitan Transportation Information," https://www.topis.seoul.go.kr/.

[11] Tmap, http://www.tmap.co.kr/.